

Gépi fordítás minőségének becslése referencia nélküli módszerrel

Yang Zijian Győző¹, Laki László^{1,2}, Prószekey Gábor^{1,2,3}

¹ Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar

² MTA–PPKE Magyar Nyelvtechnológiai Kutatócsoport

³ MorphoLogic

{yang.zijian.gyozo, laki.laszlo,
proszekey.gabor}@itk.ppke.hu

Kivonat: A gépi fordítás elterjedésével a gépi fordítók kimenetének automatikus kiértékelése is középpontba került. A hagyományos kiértékelési módszerek egyre kevésbé bizonyultak hatékonyaknak. A legfőbb probléma a hagyományos módszerekkel, hogy referenciafordítást igényelnek. A referenciafordítás előállítása idő- és költségigényes, ezért nem tudunk velük valós időben kiértékelni, és a kiértékelés minősége erősen függ a referenciafordítás minőségétől. A jelen kutatás célja, hogy olyan minőségbecslő módszert mutasson be, ami nem igényel referenciafordítást, tud valós időben kiértékelni és magasan korrelál az emberi kiértékeléssel. Az új módszer a QuEst, ami két modulból áll: tulajdonságkinyerés és modelltanítás. A tulajdonságok kinyerése során a QuEst különböző szempontok alapján minőségi mutatószámokat nyer ki a forrás- és a célnyelvi mondatokból. Majd a kinyert mutatók, illetve regressziós modell segítségével a QuEst emberi kiértékeléssel tanítja be a minőségbecslő modellt. A rendszer a betanított minőségbecslő modellel képes valós időben kiértékelni, nem használ referenciafordítást és nem utolsó sorban, mivel emberi kiértékeléssel tanított, magasan korrelál az emberi kiértékeléssel.

1 Bevezetés

Hogyan mérjük a gépi fordítás minőségét? A gépi fordítás széles körben elterjedt a hétköznapiakban. Azonban a legtöbb gépi fordító minősége megbízhatatlan. Ezért egyre több helyen merül fel igényként a gépi fordítás minőségének becslése. Elsősorban vállalati és kutatási környezetben van rá nagy szükség. Cégek esetében igen nagy segítséget tud nyújtani egy minőségi mutató a gépi fordítás utómunkáját végző szakemberek számára. Másik alkalmazása a gépi fordító rendszerek kimenetének vegyítése. Egy helyes minőségbecsléssel több gépi fordítást tudunk összehasonlítani és a jobb fordítást kiválasztva javíthatjuk a végső fordítás minőségét. Végül, de nem utolsó sorban, ismerve a fordítás minőségét ki tudjuk szűrni a használhatatlan fordításokat, illetve figyelmeztetni tudjuk a végfelhasználót a megbízhatatlan szövegrészletekre.

A gépi fordítás minőségének helyes becslése nem könnyű feladat. A hagyományos módszerek legnagyobb problémája, hogy referenciafordítást igényelnek, amelynek létrehozása igen drága és időigényes. Ezek a módszerek nem tudnak valós időben

kiértékelni és mivel ember által fordított referenciafordítás alapján értékelnek, a minőség jelentős mértékben függ a fordítás minőségétől.

A jelen kutatás ezekre a problémákra keres megoldást. A cikk egy olyan módszert mutat be, ami nem használ referenciafordítást, képes valós időben kiértékelni és magasan korrelál az emberi kiértékeléssel.

2 Gépi fordítás kiértékelő módszerek

2.1 Referenciafordítással történő kiértékelés

Kétféle módszert különböztetünk meg a gépi fordítás minőségének kiértékeléséhez: referenciafordítással történő és referenciafordítás nélküli kiértékelés.

Referenciafordítással történő kiértékelésre több módszer is rendelkezésünkre áll. A kiértékeléshez szükség van referenciamondatokra, melyeket emberek fordítottak le a forrásnyelvi korpusz alapján, majd a rendszer összehasonlítja a referenciamondatokat a gépi fordító által lefordított mondatokkal. Fontosabb referenciafordítással történő kiértékelő módszerek:

A BLEU (BiLingual Evaluation Understudy) [3] az egyik legnépszerűbb kiértékelő módszer. A BLEU azt vizsgálja, hogy a gépi fordító által lefordított mondatokban szereplő szavak és kifejezések mennyire illeszkednek pontosan a referenciafordításhoz. Az algoritmus az n -gramokból számolt értékek súlyozott átlagát adja eredményül. A módszer előnye, hogy olcsó és gyors. Hátránya, hogy nem érzékeny a szórendi átalakításokra.

Az OrthoBleu algoritmus [2] a BLEU algoritmus elméletén alapszik. A különbség, hogy amíg a BLEU szavakat kezel, addig az OrthoBleu karakterek szintjén keresi az egyezést. Ez a módszer a ragozások nyelveknél különösen előnyös, hiszen ha két szónak csak a toldaléka különbözik, a BLEU két külön szónak kezeli, és nem talál egyezést a két szó között, az OrthoBleu ezzel szemben a karakterek szintjén sokkal több egyezést talál.

A NIST (NIST Metrics for Machine Translation - MetricsMATR) [10] szintén a BLEU módszeren alapul, de pontosabb közelítést eredményez nála. Minden fordítási szegmenshez megadott módszerek alapján két független bírálatot rendelnek, majd ebből a két értékből állítják fel a végső pontszámot, amit hozzárendelnek minden fordítási szegmenshez. A NIST nem a referenciafordítást használja, hanem ezeket a bírálatok által kiszámolt pontszámokat. A NIST a szegmensekre számolt pontokból átlagot és súlyozott átlagot számol, majd ezek kombinálásával kiad egy dokumentum szintű pontszámot, ezután pedig a dokumentum szintű pontszámokkal végez rendszer-szintű kiértékelést. A NIST mérték a korreláció értéke lesz az így kapott pontszámok és a bírálatok által számolt értékek között.

A TER (Translation Edit Rate / Translation Error Rate) [8] fordítási hibaarányt számol a gépi fordítás és az emberi referenciafordítás között, az alapján, hogy mennyi javítást (szó beszúrása, törlése, eltolása, helyettesítése) kell végezni, majd a javítások számát elosztja a referenciafordítás átlagos hosszával. A TER nem kezeli a szemantikai problémákat, mert a gépi fordítás csak azt számolja ki, hogy mennyi az eltérés a

referenciafordítás és a gépi fordítás között. De közben lehet, hogy kevesebb javítással létrehozható olyan mondat, ami jelentésben megegyezik a referenciafordítással. Erre a problémára dolgozták ki a HTER (Human-targeted Translation Edit Rate / Human-targeted Translation Error Rate) módszert. A HTER módszer során célnyelvi anyanyelvű embereket kértek fel, hogy minimális lépéssel javítsák ki a gépi fordító által generált mondatokat úgy, hogy megegyezzen a jelentése a referenciamondattal. Majd az így keletkezett új referenciamondatra számolják ki a TER értéket.

2.2 Referenciafordítás nélkül történő kiértékelés

Az eddigi módszerek mind referenciafordítást igényelnek. Hátrányuk, hogy óriási emberi erőforrást igényelnek, továbbá nincsen lehetőség futási időben kiértékelni a fordítást. A referenciafordítás nélküli kiértékelő módszereket más néven minőségbecslésnek hívják. A minőségbecslés egy felügyelet nélküli automatikus kiértékelő módszer. Alapvetően statisztikai módszerekkel közelítik a problémát. A NAACL 2012 Seventh Workshop On Statistical Machine Translation keretében kiadott osztott feladatra [7] mutattak be egy teljesen újszerű, referenciafordítás nélkül történő kiértékelő módszert. Az új módszer mellett, hogy nem igényel referenciafordítást, képes futási időben kiértékelni a fordítás minőségét, továbbá a módszer segítségével minőségi mutatót adhatunk az olvasó és az utójavítást végző ember számára. A módszert a Lucia Specia által vezetett QUEST [5] és a QTLaunchPad [12] projekt keretében dolgozták ki. A két projekt közös terméke a QuEst keretrendszer [4]. A QuEst keretrendszer megvalósítja a referenciafordítás nélküli kiértékelést.

3 QuEst

A referenciafordítás nélküli kiértékeléséhez a QuEst (Quality Estimation) [6] keretrendszert kutattuk illetve használtuk fel, aminek segítségével készítettünk egy működő, referenciafordítás nélküli angol–magyar minőségbecslő és kiértékelő rendszert.

A QuEst mind a forrásnyelvi, mind a célnyelvi szövegből számtalan tulajdonságot tud kiértékelni, a nyelvfüggetlen tulajdonságoktól a nyelvspecifikus tulajdonságokig széles körben, így nem csak a fordítás pontosságára, hanem a mondat helyességére és egyéb problémákra is tud megoldást nyújtani, amelyekre más kiértékelő, mint a BLEU vagy NIST nem képesek. A QuEst keretrendszerben lévő, nyelvtől független tulajdonságok kiértékeléséhez készült funkciók felhasználhatóak a magyar fordítás kiértékelésére. A nyelvspecifikus tulajdonságok kiértékeléséhez viszont magyar nyelvre jellemző kiegészítő eszközökkel kell bővíteni a rendszert. A QuEst keretrendszer JAVA illetve Python nyelven írták, szükséges hozzá a JAVA környezet és Python osztálykönyvtárak. Két főmodulból áll: *tulajdonságkinyerő* modul és *modelltanító* modul.

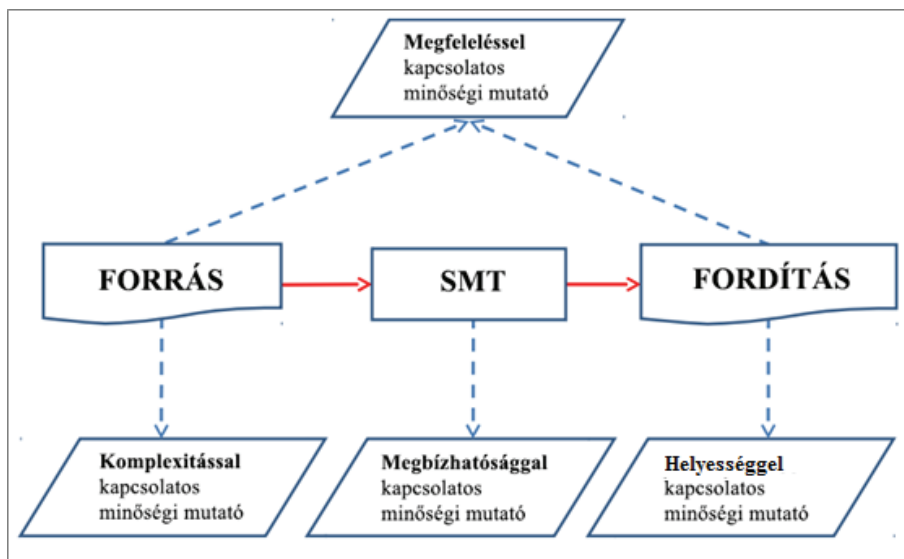
3.1 A tulajdonságok kinyerése

A tulajdonságok kinyeréséhez (Feature Extraction / Feature Sets) a QuEst-nek a forrásmondatokra illetve a gépi fordító által lefordított mondatokra van szüksége. Mivel a QuEst mondatokkal dolgozik, itt a szegmens egy mondatot jelöl. A QuEst nyelvtől függő és nyelvtől független tulajdonságokat is ki tud értékelni. A nyelvtől független tulajdonságok bármilyen nyelvre használhatóak, viszont a nyelvtől függő tulajdonságok előállításához nyelvspecifikus eszközökre is szükség van, mint például szófaji egyértelműsítő.

Az 1. ábrán azt láthatjuk, hogy a QuEst többféle típusú tulajdonságokat is ki tud értékelni: *megfelelés* (adequacy), *komplexitás* (complexity), *megbízhatóság* (confidence) és *helyesség* (fluency).

A QuEst a kinyert tulajdonságokból minőségi mutatószámokat számol minden szegmensre, így kapunk egy táblázatot, amiben a sorok az egyes szegmensek, az oszlopok a tulajdonságok.

A tulajdonságok kiértékelésére számtalan lehetőség nyílik, de nem biztos, hogy mindegyik tulajdonság releváns a minőségbecslés szempontjából. Lucia Specia kutatása [1] alapján, az angol–spanyol nyelvpárra egy 17 alaptulajdonságból (baseline) álló készletet állítottak össze. Ezek a tulajdonságok a leginkább relevánsak a minőség szempontjából. További tulajdonságok hozzáadásával nem javult jelentősen a minőség. Ebből az következik, hogy nem az a cél, hogy minél több tulajdonságot kiértékeljünk, hanem „a kevesebb néha több” elv alapján, a feladat: minőség szempontjából releváns tulajdonságokat kell keresni.



1. ábra. QuEst által kezelhető tulajdonságok típusai.

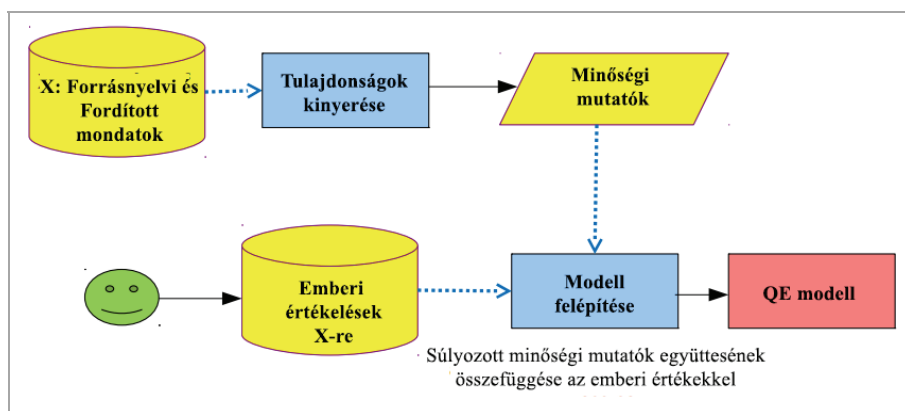
3.2 A modell felépítése

A QuEst másik főmodulja a modell felépítése, ami két részből áll: *tanulás* és *becslés*.

A tanuláshoz szükségünk van egy tanítóhalmazra. A tanítóhalmaz tartalmazza a forrásszöveget, a gépi fordító által lefordított szöveget és emberi értékeléseket. Az emberi értékelés úgy készül, hogy a gépi fordító által lefordított mondatokat emberi szakértők pontozzák két szempont alapján:

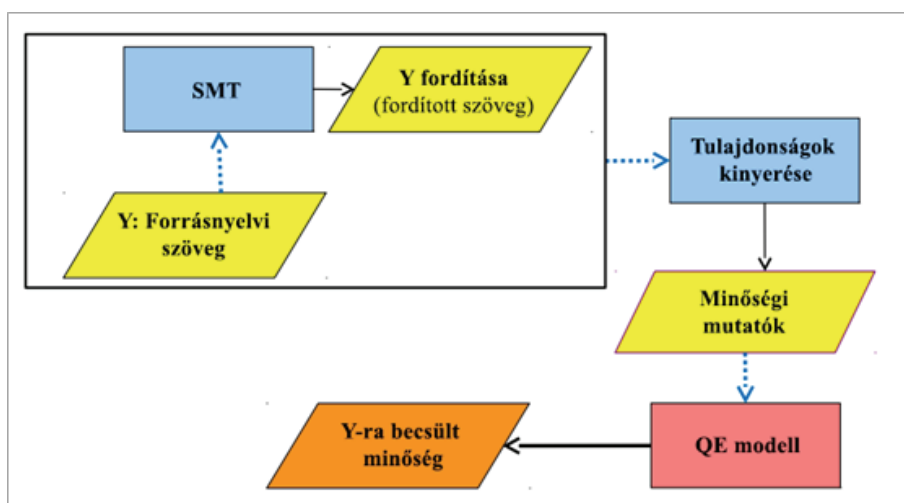
- *megfelelés* (adequacy): a lefordított célnyelvi szöveget értékeli 1–5 pontos skálán, az alapján, hogy mennyire pontos a fordítás a forrásnyelvi mondathoz képest.
- *helyesség* (fluency): a lefordított célnyelvi szöveget értékeli 1–5 pontos skálán, az alapján, hogy mennyire helyes a célnyelvi mondat.

A QuEst a tulajdonságkinyerő modell mutatószámai és az emberi értékek alapján regresszió modellel betanítja a minőségbecslő modellt. A 2. ábrán látható a tanulás folyamata.



2. ábra. Minőségbecslő modell tanításának folyamata.

A QuEst a gépi fordító által generált kimenetére – a tulajdonság kiértékelővel – mutatószámokat számol. Ezután a mutatószámokkal és az emberi értékelésekkel felépíti a kiértékeléshez szükséges modellt. Majd a tanulás során betanított modell segítségével tudja az új bemeneti mondatok minőségét megbecsülni. A minőség becslésének folyamatában már nincsen szükség emberi értékelésre. A 3. ábrán látható a becslés folyamata.



3. ábra. Minőségbecslő modell becslésének folyamata.

4 Módszerek bemutatása

A kiértékeléshez lefordított mondatokat vettünk. A forrásnyelv angol, a célnyelv magyar. A mondatokat négy különböző gépi fordítóval (Google, Bing, MetaMorpho, MOSES) lettek lefordítva, illetve a tanítóanyagban szerepel még ember által lefordított mondat is. Majd betanítottuk és kiértékeljük a QuEst keretrendszerrel. A QuEst kiértékelés minőségének mérésére a MAE (1) (Mean Absolute Error – Átlagos abszolút eltérés), RMSE (2) (Root Mean Squared Error – Átlagos négyzetes eltérés gyöke) [13] és Pearson-féle korreláció értékeket használtunk.

$$\text{MAE} = (1/N) * \sum |H(s_i) - V(s_i)| \quad (1)$$

$$\text{RMSE} = \sqrt{(1/N) * \sum (H(s_i) - V(s_i))^2} \quad (2)$$

4.1 Az emberi értékelés létrehozása

A QuEst emberek által értékelt pontszámokat használ a tanításhoz, ezért a QuEst működéséhez szükség van emberek által értékelt tanítóhalmazra. Az emberi értékelés pontszámainak létrehozásához készítettünk egy weboldalon elérhető kérdőívet¹. A kiértékeléshez önkénteseket kértünk fel, akik közép- illetve felsőfokú angoltudással

¹ <http://nlp.g.itk.ppke.hu/node/65>

rendelkeznek. A mostani eredmények 500 kiértékelt mondattal jöttek létre, de a tanítóhalmaz folyamatosan bővül.

Kettő értékelési szempontot vettünk figyelembe: *megfelelés* és *helyesség*. A megfeleléssel azt mértük, hogy a lefordított mondat tartalmilag mennyire adja vissza a forrásnyelvi mondat mondanivalóját. A helyességgel azt mértük, hogy a lefordított mondat szerkezetileg és nyelvtanilag mennyire helyes, mennyire közelít egy anyanyelvi mondathoz. A minőséget 1–5-ig terjedő skálán osztályoztuk [11] (lásd 1. táblázat).

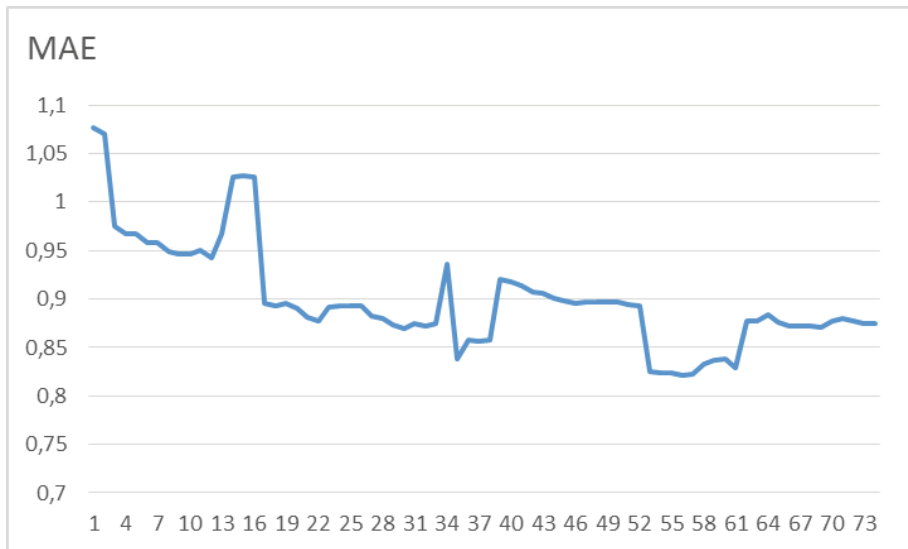
1. táblázat. Értékelési szempontok.

Megfelelés	Helyesség
0 – Nem tudom értelmezni az eredeti (angol) mondatot	
1 – egyáltalán nem jó	1 – érthetetlen a mondat
2 – jelentésben egy kicsit pontos	2 – nem helyes a mondat
3 – közepesen jó a pontosság	3 – több hibát tartalmaz a mondat
4 – jelentésben nagyrészt pontos	4 – majdnem jó a mondat
5 – jelentésben tökéletesen pontos	5 – hibátlan a mondat

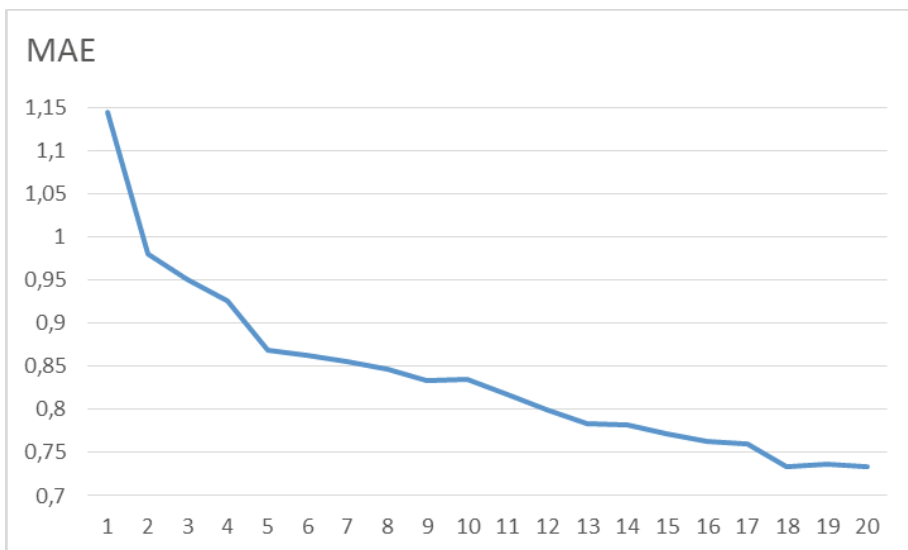
4.2 A tulajdonságok kinyerése

A tulajdonságok kinyeréséhez a QuEst keretrendszert használtunk. A Lucia Specia 2013-as cikkében [1], a QuEst kutatás során kiértékelték több mint 160 tulajdonságot, de ami igazán releváns, az csak 17 tulajdonság volt az angol–spanyol nyelvpárra. A feladat megtalálni az angol–magyar gépi fordító minőségének kiértékelése szempontjából releváns tulajdonságokat. Elsőként az angol–spanyol alaptulajdonságokkal értékeltem ki az angol–magyar mondatpárookra.

Második lépésként, kipróbáltunk további 57 tulajdonságot, majd ezekből a tulajdonságokból kivettük a nem releváns tulajdonságokat. A kiválasztás folyamata: véletlenszerűen megkevertük a 74 tulajdonságot, majd vettük az elsőt és kiértékeltük. Ezután betanítottuk a minőségbecslő modellt a kiértékelés alapján és kiszámoltuk a MAE értéket a teszhalmazra. Ezek után hozzávettük a második tulajdonságot és újra elvégeztük a kiértékelés folyamatát. Majd így tovább egy ciklussal mindig eggyel több tulajdonságot hozzávettünk és kiértékeltük (lásd 4. ábra). Ha a kiértékelés során az újonnan hozzáadott tulajdonság növelte az MAE értéket, eltávolítottuk, ha nem, akkor elvetettük. Amikor a ciklus a végére ért, előlről kezdtük a folyamatot. A ciklust elvégeztük 15-ször és a végén megvizsgáltuk, hogy melyek azok a tulajdonságok, amelyek legalább 3 alkalommal javították az eredményt. Ezeket a tulajdonságokat összegyűjtöttük, és az egész kiválasztás folyamatot előlről kezdtük a kiválasztott tulajdonsághalmazon. Így a végére maradt 20 tulajdonság, amelyekből nem tudott az algoritmus többet kizárni és ezzel a 20 tulajdonsággal sikerült elérni a legjobb MAE értéket (lásd 5. ábra).



4. ábra. Kiválasztás folyamata: 74 tulajdonsággal számoló ciklus MAE értékei.



5. ábra. Kiválasztás folyamata: 20 tulajdonsággal számoló ciklus MAE értékei.

4.3 A tanulás és a tesztelés

A gépi fordító minőségének becsléséhez a kinyert tulajdonságokkal és az emberi értékekkel a QuEst betanítja a minőségbecslő modellt. A modell teszteléséhez az 500 mondatos tanítóhalmazt bontottuk 80%-20% arányban tanító-, illetve tesztalmazra. Az így létrehozott tanítóhalmazzal betanítottuk SVR (Szupport Vektor Regresszió) [9] módszerével a kiértékelő modellt, majd a betanított modellel megbecsültük a minőségi mutatókat a tesztalmaz minden sorára. Végül a tesztalmazra kiszámolt minőségi mutatók és a tesztalmazra számolt emberi értékek alapján számoltunk MAE, RMSE és Pearson-féle korreláció értékeket.

5 Eredmények

Az optimalizáló algoritmussal egy 20 tulajdonságból álló alapkészletet állítottunk össze angol–magyar nyelvpárra. A 20 tulajdonságra angol–magyar nyelvpárra optimalizált QuEst rendszert 400 mondattal tanítottuk be és 100 mondattal teszteltük. Az alábbi táblázatban láthatóak az általunk optimalizált és javasolt 20 alaptulajdonság eredményei, összehasonlítva az angol–spanyol alaptulajdonság-készlettel kiértékelt eredményeivel, valamint a 74 tulajdonság által kapott eredményekkel.

A 2. táblázat alapján láthatjuk, hogy az angol–magyar nyelvpárra optimalizált 20 alaptulajdonság valóban jobb eredményt adott mind a 17 angol–spanyol nyelvpárra optimalizált alaptulajdonság-készlethez képest, mind a 74 alaptulajdonsághoz képest. Az eredmény alapján a QuEst a 20 alaptulajdonság készlettel körülbelül 18%-os átlag hibamértékkel tudja megközelíteni az emberi értékeket és a korreláció is elég magas (~71%). Az angol–magyar nyelvpárra optimalizált 20 alaptulajdonság-készlet a 3. táblázatban látható.

2. táblázat. Eredmények összehasonlítása.

	20 alaptulajdonság (angol–magyar)	17 alaptulajdonság (angol–spanyol)	74 tulajdonság (angol–magyar)
MAE	0,7340	0,9079	0,8746
RMSE	0,9341	1,1148	1,0573
Pearson-féle korreláció	0,7131	0,5369	0,6154

3. táblázat. A 20 alaptulajdonság angol–magyar nyelvpárra.

Tokenek száma a forrásmondatban.
Tokenek száma a célmondatban.
Átlagos tokenhossz a forrásmondatban.
Forrásmondat perplexitása.

Célmondat perplexitása.
Átlagos száma minden forrásszó fordításának a mondatban (giza küszöb: valószínűség $> 0,5$).
Átlagos száma minden forrásszó fordításának a mondatban (giza küszöb: valószínűség $> 0,2$)
Fordítások átlaga minden forrásszóra a mondatban, súlyozva a forrásnyelvi korpuszban lévő minden szó inverz gyakoriságával.
Átlagos unigram gyakoriság a második kvartilisben lévő gyakorisága (kis gyako- riságú szavak) a forrásnyelvi korpuszban.
Átlagos trigram gyakoriság a második kvartilisben lévő gyakorisága (kis gyako- riságú szavak) a forrásnyelvi korpuszban.
Forrásnyelvi korpuszban lévő negyedik kvartilisben lévő forrásszó trigramjának gyakorisága százalékban.
A korpuszban előforduló különböző trigramok százaléka.
A forrásmondatban és a célmondatban lévő kettőspontok számának különbsége abszolút értékben.
A forrásmondatban és a célmondatban lévő pontosvesszők számának különbsége abszolút értékben.
A forrásmondatban és a célmondatban lévő pontosvesszők számának különbsége abszolút értékben, célmondat hosszával normalizálva.
Írásjegyek száma a célmondatban.
Tokenek száma a forrásmondatban, amelyek nem csak a-z betűt tartalmaznak.
Forrásmondatban lévő a–z tokenek százalékának és a célmondatban lévő a–z tokenek százalékának aránya.
Igék százaléka a célmondatban.
Igék százalékának aránya a forrás és a célmondatban.

6 Összefoglalás

A kutatás során felépítettünk egy QuEst keretrendszert, és optimalizáltuk angol–magyar nyelvpárra. A kiértékeléshez szükség volt emberi értékelésekre, amihez készítettünk egy fordításiértékelő weboldalt.

Az optimalizálás során kipróbáltunk 74 tulajdonságot, amiből felállítottuk az optimalizált 20 tulajdonságból álló alapkészletet angol–magyar nyelvpárra.

A rendszer további tulajdonságok kipróbálásával tovább optimalizálható. Az általunk felépített QuEst keretrendszer megfelelő alapul szolgál a referenciafordítás nélkül történő angol–magyar gépi fordítás kiértékeléséhez és ezen a területen való további kutatásokhoz.

Hivatkozások

1. Beck, D., Shah, K., Cohn, T., Specia, L.: SHEF-Lite: When Less is More for Translation Quality Estimation. In: Proceedings of the Eighth Workshop on Statistical Machine Translation (2013) 337–342
2. FTSK, OrthoBLEU – MT Evaluation Based on Orthographic Similarities [Online] Elérhető: <http://www.fask.uni-mainz.de/user/rapp/comtrans/d05orthobleu.html>. [Hozzáférés dátuma: 2014. december 1.]
3. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual meeting of the Association for Computational Linguistics (ACL) (2002) 311–318
4. Specia, L., Shah, K., de Souza, J. G. C., Coh, T.: QuEst – A translation quality estimation framework. In: Proceedings of the 51st ACL: System Demonstrations (2013) 79–84
5. Specia, L.: QuEst – an open source tool for translation quality estimation [Online] Elérhető: <http://staffwww.dcs.shef.ac.uk/people/L.Specia/projects/quest.html> [Hozzáférés dátuma: 2014. december 1.]
6. Specia, L.: QuEst [Online] Elérhető: <http://www.quest.dcs.shef.ac.uk>. [Hozzáférés dátuma: 2014. december 1.]
7. Specia, L.: Shared Task: Quality Estimation [Online] Elérhető: <http://www.statmt.org/wmt12/quality-estimation-task.html> [Hozzáférés dátuma: 2014. december 1.]
8. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A Study of Translation Edit Rate with Targeted Human Annotation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (2006) 223–231
9. Welling, M.: Support Vector Regression. University of Toronto (2004)
10. NIST, The NIST 2008: Metrics for MACHine TRANslation” Challenge (MetricsMATR) (2008)
11. Koehn, P.: Statistical Machine Translation. 1st ed. Cambridge University Press, New York, NY, USA (2010)
12. QTLaunchPad, „QTLaunchPad,” [Online] Elérhető: <http://www.qt21.eu/launchpad> [Hozzáférés dátuma: 2014. december 1.]
13. Hyndman, R. J., Koehler, A. B.: Another look at measures of forecast accuracy (2005)